Macchine assassine oltre il Rubicone delle nostre esistenze

– Luca Celada, 04.06.2023

Guerra tech Gli avvertimenti degli stessi sviluppatori sulle incognite dell'IA addestrata per uccidere autonomamente nei conflitti del futuro

Ha fatto molto parlare recentemente l'episodio raccontato dal direttore del centro di intelligenza artificiale della Us Air Force presso l'MIT. A un congresso in Inghilterra, il colonnello Tucker "Cinco" Hamilton ha raccontato di una missione (simulata) di un drone programmato per distruggere le antiaeree nemiche.

Quando l'operatore umano ha tentato di revocare manualmente le direttive, il drone lo ha classificato come un atto di sabotaggio della missione primaria e ha "deciso" di neutralizzare l'interferenza bombardando la torre (virtuale) dei controllori umani. La frettolosa smentita dell'aeronautica non ha evitato che la storia rinforzasse il turbine di dubbi sollevato attorno ai pericoli dell'intelligenza artificiale e, in questo caso, sui sistemi cosiddetti di «autonomia letale».
I SISTEMI IN QUESTIONE sono anche detti lethal autonomous weapons systems (LAWS), armi letali che utilizzano sensori per acquisire obiettivi, e algoritmi per determinare – indipendentemente da input umani – quali colpire e distruggere. In soldoni, robot assassini, che rappresentano l'applicazione dell'intelligenza artificiale alla forza letale, uno scenario da distopia fantascientifica che, come stiamo scoprendo per molti aspetti della IA, è già dato in parte per acquisito, e, secondo il Pentagono, destinato a caratterizzare il futuro dei conflitti armati.

Nelle parole di un istruttore militare di West Point, intervistato dai giornalisti della radio pubblica americana: «Il nostro mestiere è uccidere, perché quindi non dovremmo avvalerci di ogni tecnologia che ci permetta di farlo in modo più efficiente e, possibilmente, con minori danni collaterali?».
LA PAROLA CHIAVE in questa frase è «possibilmente», dato che, come è noto, le incognite nella IA superano quelle di ogni precedente tecnologia. Nella fattispecie riguardano proprio l'impossibilità di prevedere future autocorrezioni di sistemi espressamente progettati per il progressivo autoapprendimento (machine learning) e l'opportunità di dare loro facoltà di uccidere. O, per usare le parole di Geoffrey Hinton, uno dei padri – e ora grande "pentito" – dell'intelligenza artificiale, «siamo su treno in corsa che a breve potrebbe iniziare a costruire da sé le proprie rotaie».

Se questo è vero per applicazioni "innocue" come ChatGPT, a maggior ragione le applicazioni belliche costituiscono un azzardo enorme, difficilmente articolabile in termini di semplice "efficientismo". Come ha dichiarato il colonnello Hamilton alla rivista Defence IQ, «Non si può discutere di intelligenza artificiale, machine learning e autonomia senza parlare dei risvolti etici».
L'INCOGNITA PRINCIPALE riguarda la rapidità di evoluzione di una tecnologia, la prima, che prevedibilmente saprà imparare e migliorare sé stessa, auto-generando modifiche alla propria programmazione. A questo riguardo le ipotesi comprendono un "decollo" soft, la progressione graduale verso un livello sovrumano di intelligenza o uno scatto improvviso.

Se è cioè applicabile il teorema della «rana bollita» o quello che i ricercatori denominano

«FOOM» (dall'onomatopeia fumettistica di un supereroe che spicca il volo). In entrambi i casi si profila un salto evolutivo che apre la porta teorica a noti tropi fantascientifici che anche i giganti del settore, come Hinton, considerano i paradigmi potenzialmente più adatti per visualizzarne i rischi.

IN PARTICOLARE, LE ARMI LETALI con giudizio autonomo non possono non far balenare il «teorema Skynet» – la "ribellione" anti umana delle macchine alla Terminator o Westworld. Né è solo Hinton a ritenere che sia «il momento di porsi urgentemente delle domande». A marzo più di mille addetti al settore (cominciando da Elon Musk), hanno firmato una petizione per una pausa volontaria nello sviluppo delle tecnologie "pensanti", per avere il tempo di adeguare al meglio le normative.
Potrebbe essere già tardi. Se non è già stata raggiunta la «singolarità tecnologica», il punto teorico in cui le intelligenze artificiali possono operare con "volontà" indipendente dai loro creatori, sono molti ormai gli esperti che ritengono varcata la soglia di un'inevitabilità tecnologica. E, petizione a parte, nessuno crede davvero che vi sarà uno stop allo sviluppo.

NELLA TRAIETTORIA UMANA, dai tempi della ruota, sono dopotutto inesistenti le rinunce "volontarie" a tecnologie disponibili. Oggi semmai perfino i trattati di non proliferazione nucleare sembrano in via di rottamazione. E proprio alla minaccia "esistenziale" delle bombe atomiche viene accostato il potenziale pericolo IA – non da luddisti in allarme, ma da chi conosce meglio questa tecnologia. Questa settimana un altro appello è stato firmato da 350 esperti di settore, un'unica lapidaria frase: «Mitigare il rischio di estinzione che pone l'intelligenza artificiale, dovrebbe essere una priorità globale, da considerare alla stregua dei pericoli posti da pandemie e guerra nucleare».

Gli avvertimenti sempre più urgenti lanciati da personaggi come Sam Altman, amministratore di Open AI (anche lui fra i firmatari), sono quantomeno ambigui, dato che si tratta degli stessi monopolisti di Silicon Valley che pubblicando "preventivamente" applicazioni come ChatGPT o Midjourney, hanno di fatto reso inevitabile l'attuale escalation. Resta il fatto che, a pochi mesi dalla commercializzazione dei primi sistemi generativi, la sensazione è che sia in moto una macchina dagli impatti radicali assicurati ma tutti ancora da verificare sulle nostre vite e che sia giunto il momento di «preoccuparsi seriamente» (sempre Hinton).

PER QUANTO RIGUARDA le applicazioni in grado, letteralmente, di ucciderci il concetto di "estinzione" potrebbe non essere un'iperbole. Per cominciare, è assodato che le tre superpotenze, Usa, Russia e Cina (più Israele) sono già impegnate in una corsa agli armamenti "intelligenti", che ognuno giustifica, secondo la nota logica dell'escalation, con i progressi compiuti dagli avversari. In particolare viene addotta la Cina, generalmente considerata il paese più avanzato, e quello che ha apertamente dichiarato l'obbiettivo di diventare leader mondiale entro il 2030. In fatto di robot letali saremmo dunque già in piena fase "dottor Stranamore".

D'altronde basta vedere il ruolo acquisito dalla tecnologia nel conflitto ucraino, già combattuto da droni kamikaze e missili intercettati da missili, per avere un anticipo di quello che al Pentagono considerano il futuro a cui prepararsi: la guerra di algoritmo contro algoritmo. In fatto di sistemi militari vige il top secret, ma sembrerebbe, da un rapporto dell'Onu, che almeno una prima operazione interamente "autonoma" di macchine contro combattenti in carne e ossa sia stata portata a termine nel 2020 in Libia, quando droni esplosivi forniti dalla Turchia al governo di Tripoli avrebbero attaccato «di loro volontà» le milizie orientali di Khalifa Haftar in ritirata. I quadrirotori di classe Kargu-2 prodotti dalla STM sono in grado di librarsi in volo e poi selezionare in autonomia obiettivi nemici su cui

esplodere.

DATO CHE, COME NOTO, i conflitti sono ottimi banchi sperimentali per gli ultimi prodotti dei fabbricanti di armi, il fronte ucraino non poteva che essere un gigantesco laboratorio di R&D per il complesso bellico-tecnologico. Qui sono stati consacrati i droni turchi Bayraktar e gli iraniani Shahed e l'uso, ad esempio, di droni nautici, le lance teleguidate per missioni di sabotaggio. Queste ultime sono ovviamente in dotazione anche agli eserciti cinesi e americani. L'ONR (office of naval research), l'arma tecnologica della Us Navy, sperimenta da almeno dieci anni con ""sciami marini", squadriglie di lance autonomamente governate da un'intelligenza artificiale centralizzata che valuta potenziali nemici e le modalità di risposta in scontri navali. Simili sistemi "disabitati" possono controllare elicotteri, droni o veicoli terrestri (la modalità è detta «fire, forget and find»).

Il modello moderno di combattimento era già stato efficacemente de-umanizzato dall'uso di droni americani in Iraq, Somalia, Afghanistan e altrove, dove missili Hellfire venivano sganciati su «sospetti combattenti» non più da piloti a bordo, ma da operatori in remoto, in asettiche sale di controllo su basi aeree in Nevada. La desensibilizzazione degli operatori aveva contribuito alla proliferazione delle incursioni con droni, diventate strumento privilegiato di proiezione geopolitica sotto Obama. Ma anche questo contributo umano remoto promette di diventare presto obsoleto.

DA ANNI DARPA, l'agenzia per i progetti di ricerca avanzata di difesa del Pentagono, sta lavorando a sistemi di IA «sostitutivi dell'apporto umano» per il combattimento aereo. Già nel 2020 aveva raggiunto l'obiettivo di una vittoria della IA su un avversario umano in un simulatore di volo. Lo scorso febbraio, nella base aeronautica di Edwards, nel deserto californiano, un'intelligenza artificiale ha pilotato autonomamente, per la prima volta, un vero caccia F16.

In ognuno di questi casi la rapidità matematica con cui gli elaboratori adottano decisioni che possono implicare la morte di persone sono già oltre la soglia dell'intervento e della comprensione umana. Si apre cioè la problematica di un'etica robotica che a sua volta rimanda a fiction speculative come le cardinali "leggi della robotica" postulate da Isaac Asimov o "l'invincibile" gioco di Tris giocato contro se stesso dal computer in War Games (il preconizzante film di John Badham del 1983.)
Saremmo dunque già entrati nell'era in cui, non solo affideremmo ad un robot la distinzione fra un fucile imbracciato da un combattente o il rastrello in mano a un agricoltore, ma dovremmo lasciare che siano macchine ad affrontare problemi "morali" come il «dilemma del carrello ferroviario» (classico esperimento filosofico che pone la scelta fra salvare cinque vite sacrificandone una sola).

Nella teoria della intelligenza artificiale queste problematiche cono ricondotte al concetto di "allineamento" delle macchine agli obiettivi umani. Ma questo apre a sua volta il problema di chi possa legittimamente definire gli obiettivi e delle modalità impiegate da macchine imperscrutabili per raggiungerli.

LE INTELLIGENZE GENERATIVE, come stanno dimostrando le reti neurali e i large language models dei ChatBot sono "addestrate" su un enorme volume di dati. Lo scibile umano cui attingono è quello raccolto in internet che, come è noto, comprende ogni genere di sapere spurio falso o tendenzioso. Le intelligenze artificiali non possono che rispecchiare questo oceano di dati e includere le sue imperfezioni nelle ulteriori elaborazioni (un po' come l'oceano senziente di Solaris). I casi già documentati di anomalie in Bot che hanno inventato, "allucinato" o esibito comportamenti imprevedibili sembrano confermare il

potenziale per scenari “non-allineati”.

Nel momento in cui neppure i loro creatori sanno comprendere a fondo i meccanismi con cui vengono generate immagini o testi “originali”, forse dotare sistemi di IA di missili ed esplosivi non è la migliore delle idee?
Lo ritiene ad esempio la Stop Killer Robots, una coalizione costituita nel 2013 che si batte per diffondere consapevolezza e organizzare opposizione agli armamenti senzienti. Nel 2018 anche il segretario generale dell’Onu, Antonio Guterres, ha fatto appello a tutti i paesi affinché rinuncino allo sviluppo di armi autonome definendole «moralmente ripugnanti e politicamente inaccettabili». Di fatto però, tempi e modalità di sviluppo sono rimaste in mano al complesso digitale militare e all’oligopolio delle piattaforme.

PER RICAPITOLARE, mentre si susseguono gli avvertimenti degli stessi sviluppatori, e le intelligenze sintetiche stanno già scardinando ordini simbolici e sociali (vedere lo sciopero degli sceneggiatori che a Hollywood è la prima vertenza attorno a problematiche di creatività e copyright poste dalla IA). Con l’operatività delle «macchine assassine» ci appresteremmo ad attraversare un ulteriore Rubicone senza praticamente parlarne.

L’appello di Guterres è stato raccolto da una trentina di nazioni appena, e mai preso seriamente in considerazione dalle superpotenze. Su questo sfondo i tentativi umani per contrastare il fait accompli rischiano in retrospettiva di assomigliare a scaramucce di retroguardia mentre l’intelligenza artificiale opera un’irreversibile mutazione ontologica.
In questo senso i killer robot sono solo la metafora più icastica di un momento che, come ci hanno appena ripetuto 350 saggi, potrebbe rivelarsi la possibile anticamera dell’estinzione. Se non della specie, allora almeno della cognizione e della condizione umana come l’abbiamo conosciuta.